AD_____

Award Number: DAMD17-00-1-0626

TITLE: High Throughput Analysis of the Role of Genomic Methylation in Breast Cancer by Methylation-Sensitive Comparative Genomic Hybridization

PRINCIPAL INVESTIGATOR: Anbazhagan Ramaswamy, M.D., Ph.D.

CONTRACTING ORGANIZATION: Johns Hopkins University
School of Medicine
Baltimore, Maryland 21231

REPORT DATE: September 2001

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	September 2001	Final (1 Sep 00 - 31 Aug 01)

4. TITLE AND SUBTITLE
High Throughput Analysis of the Role of Genomic Methylation in Breast Cancer by Methylation-Sensitive Comparative Genomic Hybridization

5. FUNDING NUMBERS
DAMD17-00-1-0626

6. AUTHOR(S)
Anbazhagan Ramaswamy, M.D., Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Johns Hopkins University School of Medicine
Baltimore, Maryland 21231

E-Mail:

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)***

We developed software for the identification and analysis of CpG island in the genomic DNA sequence. We choose a simple and efficient platform, "Microsoft Excel", which is commonly available with most of the academic scientists. Upon execution of the program, a customized workbook analyzes an entered DNA sequence for the total number and percentage cytosine and guanine nucleotides, the total number and percentage of CpG sites, and a CpG:GpC ratio. The program also displays the distribution of CpG sites in a visual format as well as in two different graphical formats. Finally, the program also displays methylation-dependent effects of bisulfite treatment on DNA sequences. Subsequently, we modified this program later, so that thousands of sequences can be analyzed at the same time. Using this modified program we have analyzed 40000 clones from the Research Genetics and identified 162 clones, which showed features consistent with 'CpG' island. By employing restriction sensitive comparative genomic hybridization we have identified 7 clones, which are probably differentially methylated in breast cancers. These clones are being evaluated for their usefulness as diagnostic markers for breast cancer.

14. SUBJECT TERMS	15. NUMBER OF PAGES
breast cancer, methylation, software, hybridization	10
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Table of Contents

Introduction

Cancer results from abnormalities in critical genes that regulate normal cellular growth and development. These abnormalities arise in two classes of interacting genes: those that facilitate cell growth and tumor formation, whose over expression induces cancer, and those that inhibit these process whose loss or mutation causes cancer. Genes that belong to the second category are called tumor suppressor genes. These tumor suppressor genes can be inactivated by two different mechanisms. One mechanism involves the mutation of the coding sequences within the gene. The other mechanism, which is being increasingly appreciated more and more, is the inactivation of the tumor suppressor genes by promoter methylation. These promoters are portions of DNA sequences found closed to the starting portion of the gene and often undergo changes in their molecular structure (methylation) thereby preventing effective transcription of these genes. This project aims at developing software tools to find out these areas which are rich is CpG

Body

Our first task was to develop the software for the identification of CpG islands in genomic DNA sequences. While this project was started there was no software available which could precisely do this function. So we set out to achieve this goal by developing a software exclusive for this purpose. We decided to choose a simple and efficient platform such "Microsoft Excel" which is commonly available with most of the academic scientists. This spread-sheet based program is very versatile for handling large DNA sequences and also can be easily programmed using "Visual Basic for Applications" to

develop software for custom use. We initially developed a custom Excel spread-sheet in which a sequence to be analyzed could be copied and pasted from other public domain databases (Anbazhagan et al, 1991). Upon execution of the program, a customized workbook analyzes an entered DNA sequence for the total number and percentage cytosine and guanine nucleotides, the total number and percentage of CpG sites, and a CpG:GpC ratio. The program also displays the distribution of CpG sites in a visual format as well as in two different graphical formats. Finally, the program assists in laboratory studies of DNA methylation that employ bisulfite modification of DNA by displaying methylation-dependent effects of bisulfite treatment on DNA sequences.

While this program is very useful for the analysis of individual sequences, it couldn't handle large number of sequences for large-scale analysis. So we have modified this program later, so that thousands of sequences can be analyzed at the same time. Using this modified program we have analyzed 40000 clones from the Research Genetics and identified 162 clones, which showed features consistent with CpG island. By employing restriction sensitive comparative genomic hybridization we have identified 7 clones, which are probably differentially methylated in breast cancers. These clones are being evaluated for their usefulness as diagnostic markers for breast cancer.

Key Research Accomplishments

Developed a software tools to analyze the genomic DNA and cDNA and identify specific regions, which are rich in CG content. Using this program identified 162 clones from which show features suggestive of being potential sites for CpG methylation.

Employing high through put technology for the identification of differentially methylated clones in breast cancer, selected 7 clones, which show differential methylation in breast cancer.

Reportable Outcomes

Manuscript published in BioTechniques journal regarding software developed for the analysis of CpG islands (Anbazhagan et al, 1991).

Conclusions

While the software developed for the identification of CpG islands will be very useful in future work in this area, further evaluation of the differentially methylated clones would help to develop new breast cancer markers for diagnostic use.

References

Anbazhagan R, Herman J, Enika K, Gabrielson E. Spreadsheet-based program for the analysis of DNA methylation. BioTechniques 2001, 30:110-4.

Appendices:

1. Paper published in BioTechniques regarding software developed for the analysis of CpG islands.

Spreadsheet-Based Program for the Analysis of DNA Methylation

BioTechniques 30:110-114 (January 2001)

Ramaswamy Anbazhagan, James G. Herman, Kalyan Enika, and Edward Gabrielson
The Johns Hopkins University School of Medicine, Baltimore, MD, USA

ABSTRACT

Methylation of DNA in CpG dense regions of gene promoters (CpG islands) is important for transcriptional inactivation of selective genes in normal and neoplastic cells. Here, we present a spreadsheet-based program adapted from Microsoft® Excel® that is useful for identifying CpG islands and for assisting in the laboratory analysis of DNA methylation of these regions. Upon execution of the program, a customized workbook analyzes an entered DNA sequence for the total number and percentage cytosine and guanine nucleotides, the total number and percentage of CpG sites, and a CpG:GpC ratio. The program also displays the distribution of CpG sites in a visual format as well as in two different graphical formats. Finally, the program assists in laboratory studies of DNA methylation that employ bisulfite modification of DNA by displaying methylation-dependent effects of bisulfite treatment on DNA sequences.

INTRODUCTION

While the CpG dinucleotide occurs at less than 10% of its expected frequency in the human genome, many gene promoter regions have high densities of CpG sites. These regions are known as CpG islands, and methylation of cytosines in CpG islands is a common mechanism for transcriptional inactivation of genes on the inactivated X-chromosome in females and of silenced imprinted alleles on autosomal chromosomes. Furthermore, methylation is increasingly recognized as an important mechanism for inactivation of tumor suppressor genes in neoplasia (reviewed by Baylin et al. in Reference 2).

CpG islands are defined as regions of DNA ranging in size from 0.5 to 2 kb that have a C + G content of greater than 60% and a ratio of CpG to GpC of at least 0.6 (3,4). Identification of CpG islands is a critical initial step for studying transcriptional regulation of genes by methylation. The status of CpG islands can then be evaluated in the laboratory using restriction enzymes that are methylation specific, by sequencing bisulfite modified DNA, or by methylation-specific PCR (MSP) (5) of bisulfite-modified DNA. Here, we report the adaptation of a commonly available spreadsheet program, Microsoft® Excel® 2000, as a tool for identifying CpG islands and assisting in the analysis of DNA methylation.

MATERIALS AND METHODS

Downloading and Opening the Program

We have created an Excel file with a built-in program that performs all major functions required for the analysis of CpG islands. The program works with Excel 2000. The program file (CpG-Win) can be downloaded from the Software Library on the Internet at http://128.20.85.49/genomics. After downloading the file, the program can be started in two ways: *(i)* either open the downloaded file or *(ii)* start the Excel program and open the downloaded file from within the program using File/Open. A dialogue box appears, asking to enable or disable macros. Click on the Enable Macros button. When the program opens, you will notice that a new menu named Analyze-CpG is added to the program menu bar. This menu has seven menu items, namely New-Workbook, Format, Count-CpG, Mark-CpG, Bis-Modify, and Make-Graph. These menu items can be executed to perform various functions necessary for the analysis of CpG islands within the DNA sequences.

Opening a New Workbook and Entering a Sequence

The Excel workbook is organized as a collection of sheets, with each sheet's name appearing on a tab at the bottom of the workbook. Navigation between sheets is possible by clicking on the tab at the bottom. Refer to the Microsoft User's Guide for further explanation about the organization and use of Excel workbook and sheets. To use this program, go to the newly added Analyze-CpG menu and click on the New-Workbook menu item. This command will open a new workbook with four sheets. The program automatically formats the Sheet2 as shown in Figure 1. To analyze a specific DNA sequence, enter the nucleotide sequence in cell A1 of Sheet1. Nucleotides in the sequence

may be entered in either uppercase or lowercase. Ambiguous nucleotides may be designated as "n". The cells can hold more than 30 000 characters and can therefore accommodate long sequences. These sequences may be copied and pasted from other documents. GenBank® sequences viewed through Internet browsers can also be directly copied and pasted along with numbers and spaces. If you copy and paste sequences from GenBank, you will notice that the sequence occupies multiple rows in Sheet1. After entering or pasting the sequence in cell A1 of Sheet1, go to the Analyze-CpG menu and click on the Format menu item. This command removes all the numbers and spaces from the sequence (if there are any), concatenates the sequences in all the rows, and also transfers the sequence from cell A1 of sheet1 to cell J2 of Sheet2. The command also activates Sheet2 so that you can see the transferred sequence there.

The program automatically fills up cell J3 of Sheet2 with a formula that will count and display the total number of nucleotides in cell J2 of Sheet2. When a sequence is transferred from Sheet1 to cell J2 of Sheet2, the nucleotide count of that sequence is automatically displayed in cell J3. The nucleotide count in cell J3 is also updated automatically whenever any sequence entry is altered in cell J2.

Analyzing the CpG and GpC Distribution in Different Regions of the Sequence

The next step is to designate which portions of the sequence are to be analyzed for CpG islands. This is done by entering the starting and ending position numbers of the fragments to be analyzed in column G and H, respectively, below the labels Start and End. Up to six fragments can be entered in these columns with the starting and ending position numbers of their nucleotides. After entering these values, go to the Analyze-CpG menu and click on the Count-CpG menu item. After a few seconds delay (depending on the length of the sequence), the analyzed data are displayed in columns J–X as shown in Figure 2.

Column J (labeled "Length") displays the length (number of nucleotides) of each of the fragments entered in columns G and H. Column L (labeled "Number of C+G") displays the total number of cytosine and guanine nucleotides within each of the fragments. The percentage of cytosine and guanine nucleotides within each fragment is displayed in column N (labeled "(C+G)%"). Column P (labeled "Number of CpG") displays the total number of CpG islands within each of these fragments. The percentage of CpG islands within each of these fragments is displayed in column R (labeled "CpG%"). Similarly, the total number and the percentage of GpC base pairs are displayed in columns T and V (labeled "Number of GpC" and "GpC%", respectively). The last column, X, shows the CpG:GpC ratio in the sequence.

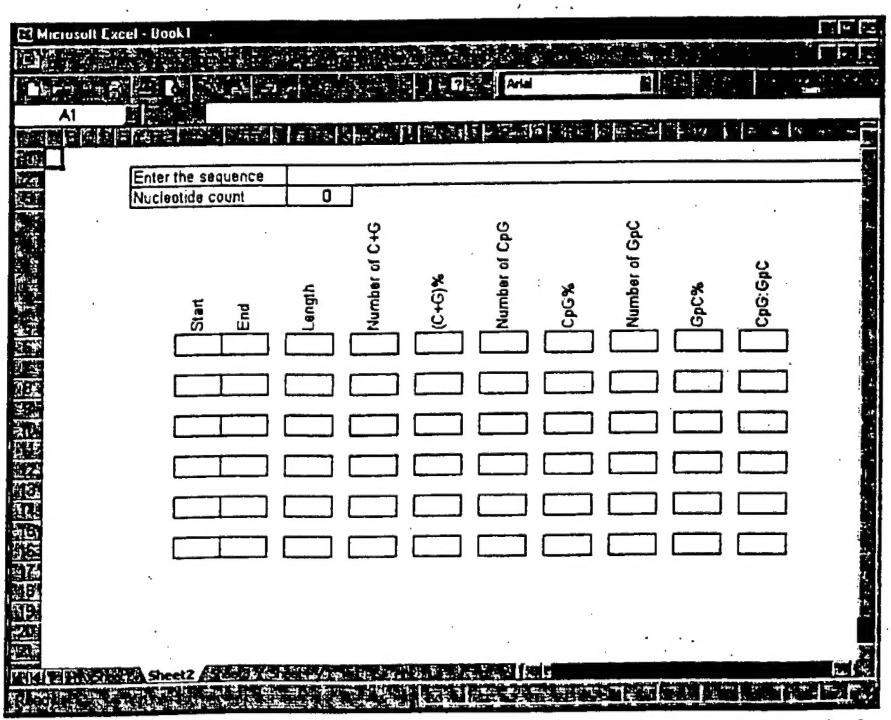

Figure 1. Sheet2 of a new workbook opened by executing the menu item New-Workbook in the CpG-Analyze menu.

Marking the CpG and GpC Base Pairs for Visual Display

Although the various parameters of the CpG islands are analyzed as described above, it does not give a visual impression of the distribution of the CpG and GpC sites within a sequence. The next menu item, Mark-CpG, is meant to perform this function. When this menu item is executed, the program takes the sequence in cell J2 of Sheet2 and displays it in columns B and D (labeled "CpG Marked" and "GpC Marked", respectively) of Sheet3 in a vertical column format with each nucleotide occupying a single cell. The program also highlights the CpG sites in column B and GpC sites in column D with a bright pink background. This highlighting helps to visually appreciate the distribution of the CpG and GpC sites (Figure 3). This visual effect can be better appreciated by clicking on the View menu bar and executing the Zoom function, which are built-in Excel commands. By adjusting the zoom ratio to 10% of the original, an overall view the marked sequence can be seen.

Displaying the Bisulfite-Modified DNA Sequence

MSP and sequencing of bisulfite-modified DNA are commonly employed to analyze methylation of DNA. These techniques both involve treating the DNA with bisulfite, which chemically modifies the unmethylated cytosine residues to uracil. In effect, this modification is reflected as thymine after sequencing or MSP. Bisulfite treatment of DNA therefore results in different prod-

ucts, depending on the methylation status of the DNA before modification.

The next command in our custom menu, Bis-Modify, helps to identify possible nucleotide sequences after bisulfite treatment, depending on prior methylation status. When this menu item is executed, the DNA sequence from cell J2 of Sheet2 is displayed in columns G, H, and I of Sheet3 (labeled "Original", "Unmethylated" and "Methylated", respectively), in a vertical format, with each nucleotide occupying a single cell. The column labeled "Original" displays the original sequence for the purpose of comparison with others. The column labeled "Unmethylated" displays the modified sequence that would result from bisulfite treatment of unmethylated DNA. Note that in this situation, all cytosine nucleotides would be converted to thymine regardless of whether or not they are associated with a part of a CpG island. These modified nucleotides are highlighted with a bright pink background. The next column labeled "Methylated" displays the sequence that would result from bisulfite treatment of DNA with methylated sequences in CpG islands. In this situation, all cytosines not followed by guanine (thus not making a CpG pair) are modified to thymine, while those followed by guanine (thus making a CpG pair) are retained as cytosines. As in the previous case, all the modified nucleotides are highlighted with a bright pink background. Thus, the execution of this menu item displays all three sequences (original, unmethylated, and methylated) side by side with the modified nucleotides being highlighted. This display format facilitates comparison of the sequences and identification of possible sites of methylation. The results of bisulfite modification of the complementary strand are displayed in columns K, L, and M.

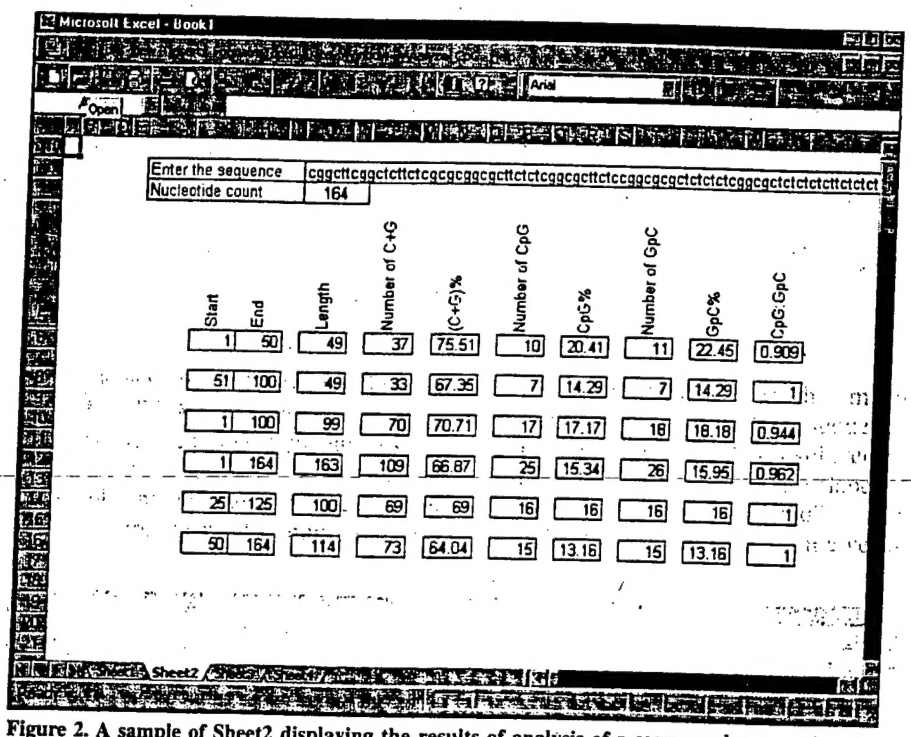

Figure 2. A sample of Sheet2 displaying the results of analysis of a sequence by executing the Count-CpG menu item.

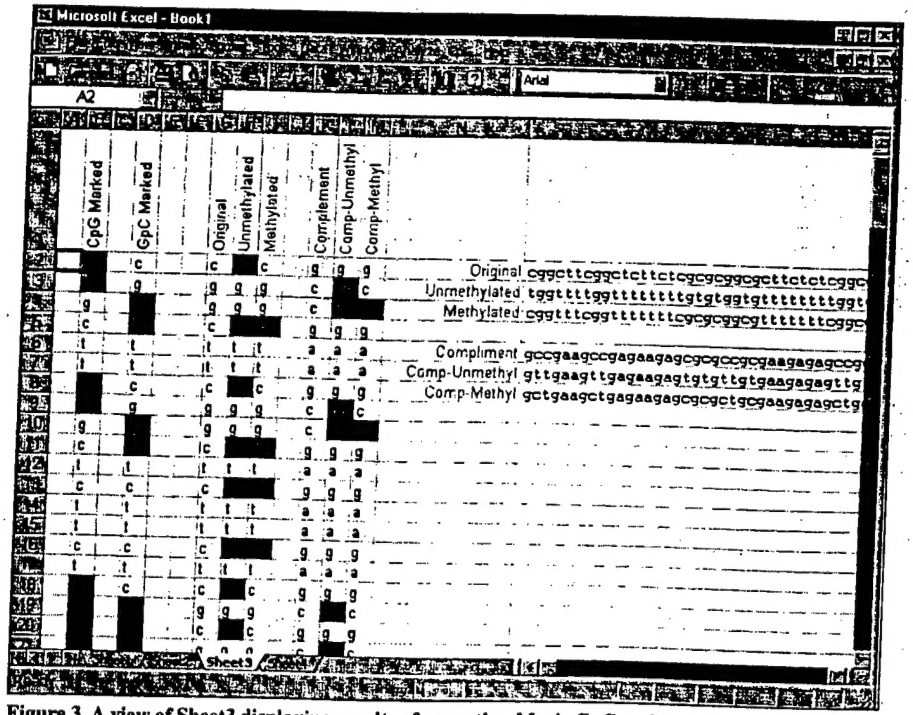

Figure 3. A view of Sheet3 displaying results of executing Mark-CpG and Bis-Modify menu items.

To design primers for MSP, it may be desirable to copy these unmodified and modified sequences to another file or to another program used for designing primers. To make it easy to copy these sequences, they are also displayed in a horizontal format with each sequence occupying a single cell as shown in Figure 3. Actually, the menu item Bis-Modify performs this function also. The unmethylated or methylated version of DNA sequence could be easily copied from column O.

Displaying the CpG and GpC Base Pairs in a Graphical and Pseudographical Formats

The next menu item, Make-Graph, is used to display the distribution of CpG and GpC base pairs in graphical formats. When this command is execut-

ed, the program takes the sequence in cell J2 of Sheet2, analyzes the CpG distribution, and plots two graphs. First, the program scans the sequence through a window of 10 bp, computes the CpG frequency, makes a line graph, and displays it in Sheet4 as shown in Figure 4. The Y-axis in this graph represents the exact number of CpG base pairs counted within a window of 10 bp. Then, the program scans the sequence again through a window of 100 bp, computes the CpG frequency, and displays the graph in Sheet4 just below the earlier graph. The Y-axis in this graph represents the percentage of CpG frequency and has a fixed range of 0%–12%. While the first graph will be useful for the analysis of shorter fragments, the second will be very useful to analyze larger fragments.

The program also makes a text string "pseudograph" for CpG and GpC base pairs. In this pseudograph, the CpG or GpC pairs are represented by the "|" symbol, and all other possible pairs of nucleotides are represented by "." symbol. For example, a sequence such as "acgtcgac" is displayed ".|.|.." to represent the distribution of the CpG pairs. These pseudographs are displayed just above the previous graph with labels "CpG base pairs" and "GpC base pairs". This gives another visual tool to appreciate the distribution of the CpG islands in a sequence. The sequence and the results can be saved using the built-in Save As command in Excel. The user can then open a new workbook and continue analyzing new sequences, if necessary.

DISCUSSION

One important feature that we have not included in this program is the ability to search for restriction sites, especially for methylation-specific enzymes within the DNA sequence. We did not include this feature in the current program because our earlier programs, Win-Align and Mac-Align, can perform these functions effectively (1). In summary, we have described a spreadsheet-based program for the analysis of CpG islands. This program has features to copy and paste sequences from the GenBank database and would be very useful for identifying CpG frequency in different regions of DNA. The user can get a numerical output of the CpG frequency in selected regions and appreciate the output visually through a color-shaded vertical display and through graphical and pseudographical formats. The program also outputs the sequence that is expected after bisulfite treatment of DNA. This feature will be especially useful for designing primers for methylation-specific PCR.

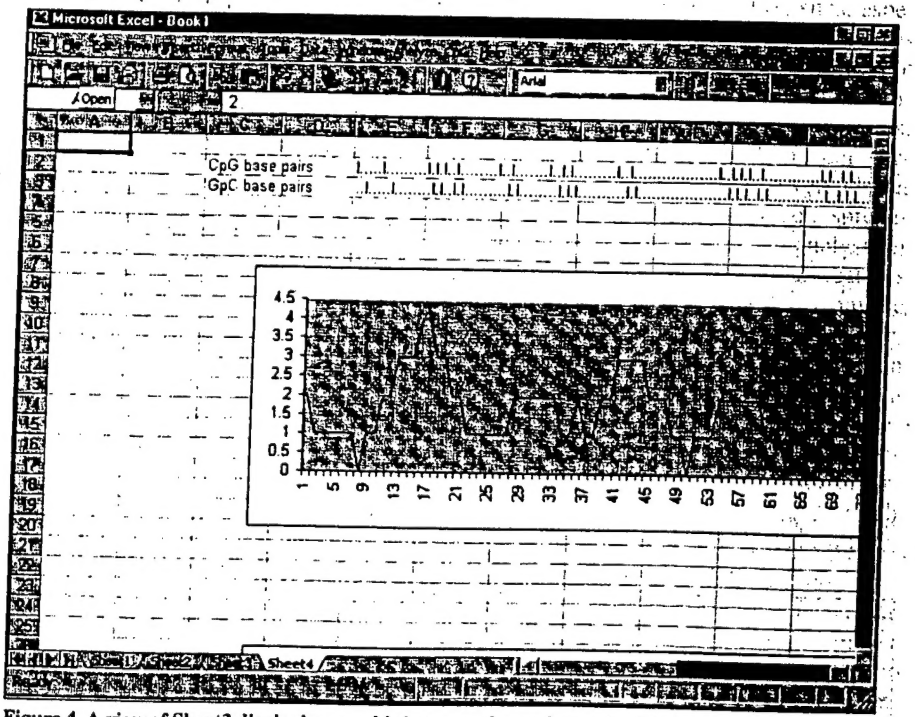

Figure 4. A view of Sheet3 displaying graphic image and pseudograph of CpG and GpC base pairs.

REFERENCES

1. **Anbazhagan, R. and E. Gabrielson.** 1999. Spreadsheet-based program for alignment of overlapping DNA sequences. BioTechniques *26*:1180-1185.
2. **Baylin, S.B., J.G. Herman, J.R. Graff, P.M. Vertino, and J.P. Issa.** 1998. Alterations in DNA methylation: a fundamental aspect of neoplasia. Adv. Cancer. Res. *72*:141-196.
3. **Bird, A.P.** 1986. CpG-rich islands and the function of DNA methylation. Nature *321*:209-213.
4. **Gardiner-Garden, M. and M. Frommer.** 1987. CpG islands in vertebrate genomes. J. Mol. Biol. *196*:261-282.
5. **Herman, J.G., J.R. Graff, S. Myohanen, B.D. Nelkin, and S.B. Baylin.** 1996. Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands. Proc. Natl. Acad. Sci. USA *93*:9821-9826.

Received 10 April 2000; accepted 27 September 2000.

Address correspondence to:
Dr. Ramaswamy Anbazhagan
Department of Pathology
The Johns Hopkins University School of Medicine
Room 301
418 North Bond Street
Baltimore, MD 21231, USA
e-mail: anba@jhmi.edu